**Predicting MLS Results**

Alexandre Moreira Batista

300344513

Armando Defendi Rossi

3003340211

CSIS 4495-001

Stephen Chiong

Douglas College

**Predicting MLS results**



**Executive Summary:**

This project aims to give sports fans accurate predictions for Major League Soccer (MLS) games using robust machine-learning models. By utilizing web scraping, database management, machine learning analysis, and an Android application, this project will deliver valuable predictions to users without needing them to pay exorbitant fees to sports betting companies. This project will cover all aspects of the Data Analytics program at Douglas College. The primary beneficiaries will be soccer fans with access to accurate game predictions at no cost.

**Table of Contents**

[1. Introduction 4](#_Toc129540428)

[A. Statement of the Problem 4](#_Toc129540429)

[B. Significance of the Study 4](#_Toc129540430)

[C. Technologies 4](#_Toc129540431)

[D. Project Implementation and timelines 4](#_Toc129540432)

[2. Actual Report – Progress Report 2 5](#_Toc129540433)

[A. Project Implementation 5](#_Toc129540434)

[i. Complete Steps 5](#_Toc129540435)

[a) Development and testing of database architecture to receive data. 5](#_Toc129540436)

[b) GitHub configuration for Project 7](#_Toc129540437)

[c) Development of Python scripts for web scraping 9](#_Toc129540438)

[d) Stage 01 - Development of the Android App 10](#_Toc129540439)

[e) Web scrapping for Machine Learning 11](#_Toc129540440)

[f) Plots and Descriptive Analysis 15](#_Toc129540441)

[g) Data Cleaning and Preparation 18](#_Toc129540442)

[3. Software Design Architecture 20](#_Toc129540443)

[4. References 22](#_Toc129540444)

# Introduction

## Statement of the Problem

The sports betting industry has seen significant growth recently, with large companies and start-ups competing for customers and revenue. While many companies offer predictions for game results, access to this valuable information often comes at a cost. This project aims to provide a cost-effective solution for all sports fans to access accurate predictions for MLS games.

## Significance of the Study

This project aims to deliver predictions that improve the odds ratio when determining the winner of a game. By utilizing techniques such as web scraping, database management, machine learning, and analytics, this project will also identify features that impact the accuracy of predictions. Additionally, this project will deliver an Android application, providing users with easy access to past and future game predictions.

## Technologies

The project will make use of the following technologies:

* Programming Language: Python for web scraping and data prediction.
* Database: MySQL
* Front-end and Backend: Android Studio

## Project Implementation and timelines

* Development and testing of database architecture to receive data - 1 week.
* Development and testing of Python scripts for web scraping - 2 weeks.
* Research on modelling and machine learning uses in soccer results prediction - 2 weeks.
* Preparation and setup of a virtual machine in the cloud (Azure) to run the database and modelling - 1 week.
* Data wrangling in Python - 1 week.
* Data modelling (machine learning to predict and check prediction accuracy) - 2 weeks.
* Development of the Android App - 4 weeks.
* Testing the entire ecosystem and fixing any bugs - 1 week.

# Actual Report – Progress Report 2

## Project Implementation

### Complete Steps

#### Development and testing of database architecture to receive data.

Following experiments conducted on Azure SQL (as shown in figure 01), our team opted to proceed with MySQL instead, as depicted in figure 02. This solution is hosted online and can be accessed via Python scripts (for web scraping) and our Android application.

As we were working on developing an Android application and collecting information from the MLS website, we noticed that the website was constantly being revamped. This posed a major obstacle for us while using web scraping techniques to retrieve data from the site. Since the MLS website uses JavaScript to construct its pages, modifications occur rapidly. Consequently, we decided to switch to ESPN's website, which provides similar data. We made this change when we needed to refresh our database with the latest scores of recent matches, and now the scores are updated one day after the matches are played.

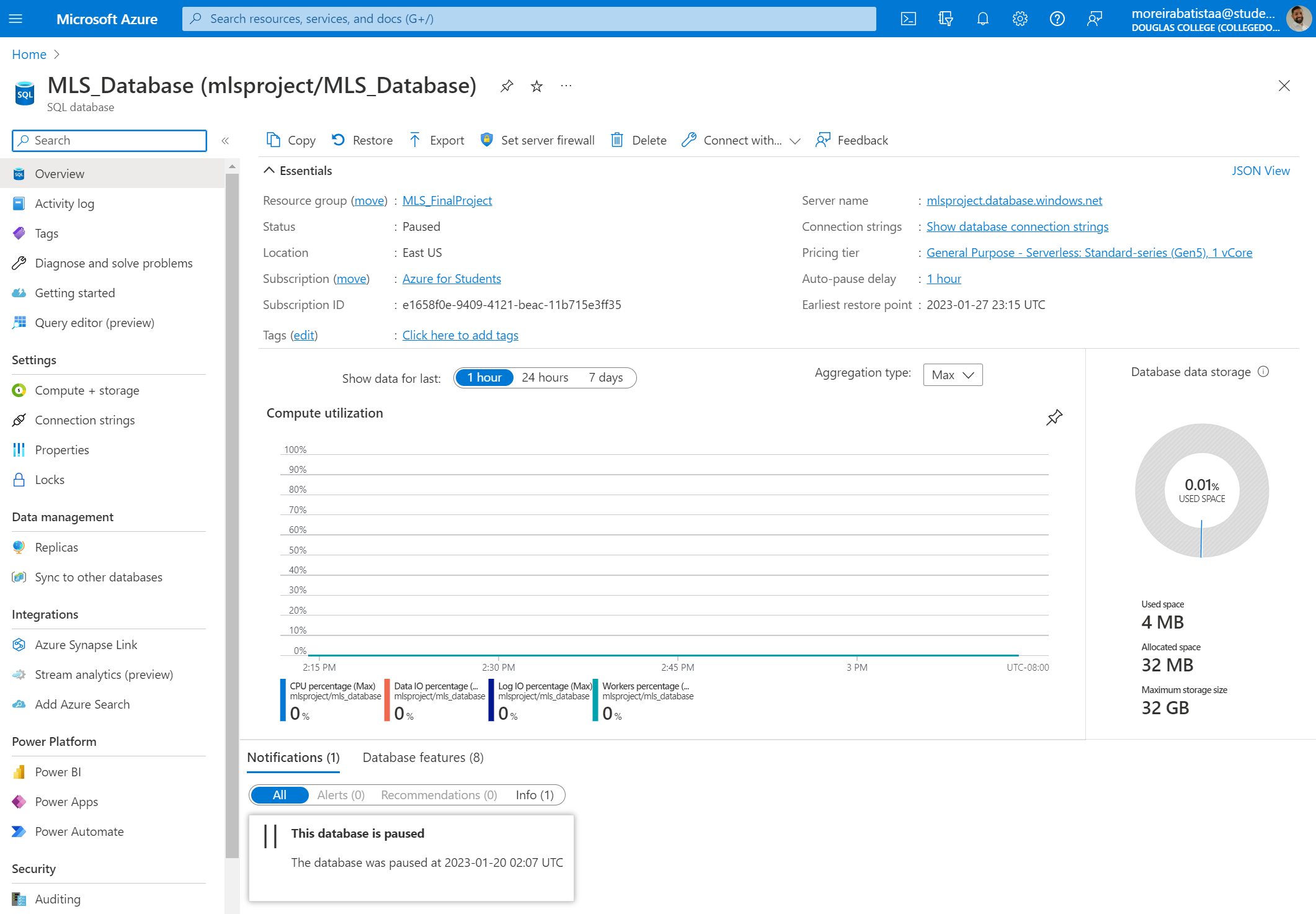


Figure 1 - Azure SQL Server

**Graphical user interface, text, application

Description automatically generated**

Figure 2 - MySQL structure

The first tables were created, and the connection tested.

Database configuration: Alexandre on 16-20/Jan/2023  
Database creation of tables/fields: Armando on 16-20/Jan/2023

#### GitHub configuration for Project

A repository was created to manage the versioning and save all the scripts and codes online as shown on figure 03.

GitHub creation: Armando on 16-20/Jan/2023

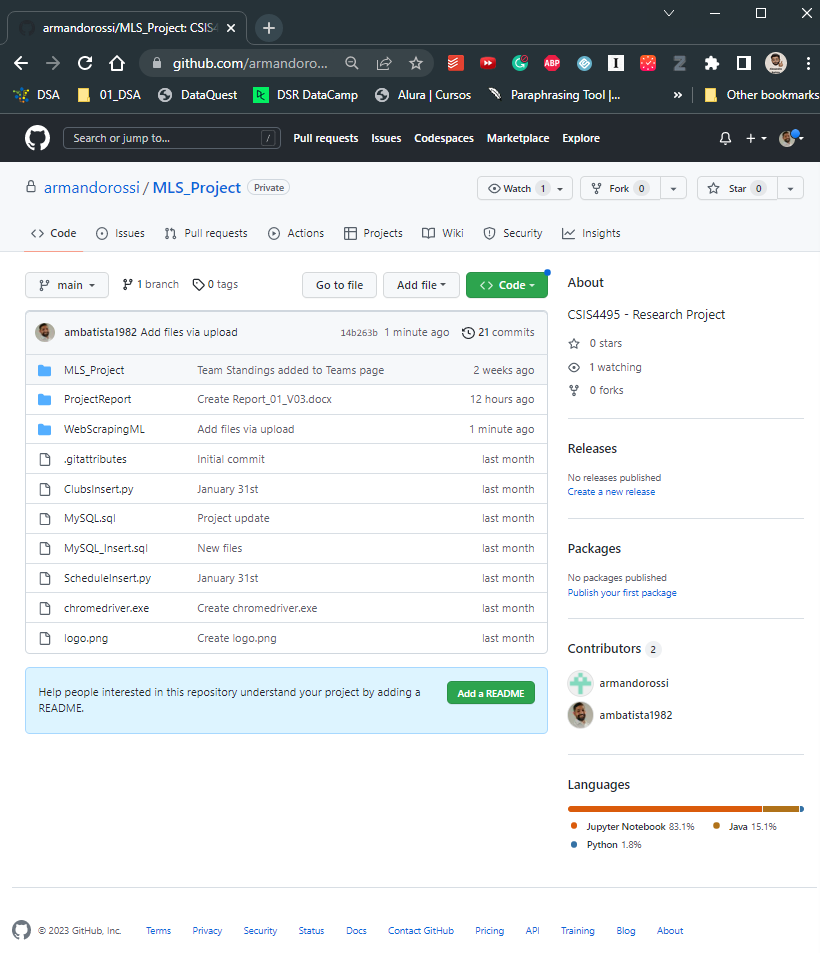


Figure 3 - GitHub directory

#### Development of Python scripts for web scraping

The scripts were developed using some libraries from Python based on the page: [*https://www.mlssoccer.com/schedule/scores#competition=mls-regular-season*](https://www.mlssoccer.com/schedule/scores#competition=mls-regular-season). It is a dynamic page, and the usual libraries, such as Request and BeautifulSoap, do not work with this kind of page.

Graphical user interface, text, application, email

Description automatically generated

Figure 4 - Python script

The scripts bring data from 2015 to actual year and insert it into the schedules table. We got 3,573 entries for that period.

Graphical user interface, application

Description automatically generated

Figure 5 - Schedule table details

Python scripts: Armando on 23-26/Jan/2023

#### Stage 01 - Development of the Android App

The first visuals were developed, including the Login, Register, Schedule, and User administration page layout.  
Graphical user interface, text, application, email

Description automatically generated Graphical user interface, text, application, email

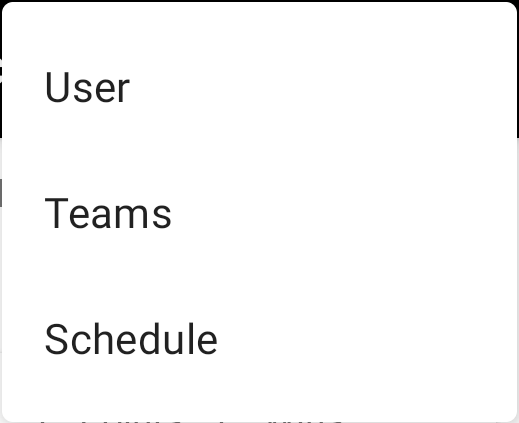
Description automatically generated  
*Figure 6 - Login page* *Figure 7 - Registration page*

Graphical user interface, application

Description automatically generated Graphical user interface, application

Description automatically generated  
*Figure 8 - User page* *Figure 9 - Schedule page*

To accommodate the upcoming modifications, the application has undergone a revamp. A fresh menu has been introduced to enable users to switch between the User (admin access only), Schedule, and the newly developed Team screen. The Team screen showcases a comprehensive list of all MLS teams and their standings. It also allows users to choose a specific year to review each team's standings.

 Graphical user interface, application

Description automatically generated

Figure 10 – Menu Figure 11 – Teams page

Graphical user interface, text, application

Description automatically generated  
*Figure 12 – GitHub repository*

App Development: Armando on 23/Jan to 13/Mar/2023

#### Web scrapping for Machine Learning

In order to improve the accuracy of the machine learning model's forecasts, it was determined that additional data was needed beyond the results of each game. To acquire this information, the development of Python scripts to scrape data from the FB Ref website was initiated.

The struture of the Machine Learning Scripts was create using the following steps:

1. Import the additional Libs
2. Defining the standard Site Adresses
3. Creating Function to Extract the Data from the site
4. Creating SQL Functions to Query, Insert and Truncate tables
5. Function GetInsert (grouping toguether all Functions)
6. Testing Functions
7. Get and Inserting Data from web to Database (2015-2023)

*Graphical user interface, text, application

Description automatically generated*  
*Figure 13 – Python web scrapping for ML*

Text

Description automatically generated  
*Figure 14 – Python web scrapping for ML*

Graphical user interface, text, application

Description automatically generated

*Figure 15 – Python web scrapping for ML*

Web scraping Development: Alexandre on 23/Jan to 26/Feb/2023

#### Plots and Descriptive Analysis

Graphical user interface, text

Description automatically generated with medium confidence

A picture containing text

Description automatically generated

Graphical user interface, application, table

Description automatically generated

Chart, line chart

Description automatically generated

#### Data Cleaning and Preparation

Text

Description automatically generated

A picture containing calendar

Description automatically generated

Table

Description automatically generated

Text

Description automatically generated

Data analysis and cleaning: Alexandre on 26/Feb to 12/Mar/2023

Next steps:

* 11. Modeling data to features engineering
* 12. Creating Machine Learning Models
* 13. Measure accuracies, create Predictions and export to Mobile Applications.

# Software Design Architecture

**1. Front-end**

* Android application for users to access the sports predictions and other features.
* User interface design to allow easy navigation and interaction.

**2. Back end**

* Web scraping module to collect data on MLS games, including past results and upcoming fixtures.
* Machine learning module to analyze the data and generate accurate predictions for future games.
* Database management module to store and manage the data collected from web scraping and machine learning modules.
* Analytics module to identify the features that impact the accuracy of predictions.
* API layer to expose the prediction data and other features to the Android application.

**3. Integration and Deployment**

* Integration of the front-end and back-end modules to create a seamless user experience.
* Deployment of the application on cloud platforms or other hosting environments to ensure scalability and availability.

The software design architecture above is based on a client-server architectural style, with the Android application as the client and the back-end modules as the server. The system will utilize web scraping, machine learning, and database management techniques to collect and analyze data and generate accurate predictions for MLS games.

Additionally, the analytics module will identify the features that impact the accuracy of predictions. The Android application will provide easy access to past and future game predictions and other features. Finally, the system will be deployed on cloud platforms or other hosting environments for scalability and availability.

# References

Major Soccer League. Schedule & Scores: <https://www.mlssoccer.com/schedule>

FB Ref – Soccer References Databases: <https://fbref.com/en/comps/22/history/Major-League-Soccer-Seasons>

ESPN. Schedule & Scores: <https://www.espn.com/soccer/schedule/_/league/usa.1>